

# Image Transformation for Object Tracking in High-Resolution Video

Tae Eun Choe, Krishnan Ramnath, Mun Wai Lee, Niels Haering  
*ObjectVideo Inc.*  
{tchoe, kramnath, mlee, nhaering}@objectvideo.com

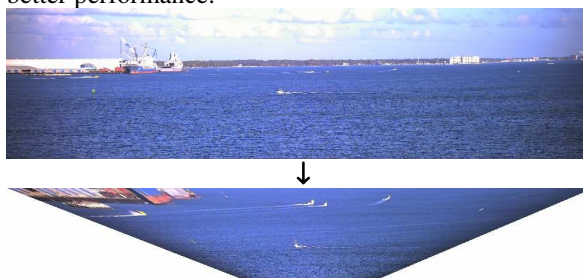
## Abstract

We propose a new method for warping high-resolution images to efficiently track objects on the ground plane in real time. Recently, the emergence of high resolution video cameras (>5 megapixels) has enabled surveillance over a much larger area using only a single camera. However, real-time processing of high resolution video for automatic detection and tracking of multiple targets is a challenge. When the surveillance camera covers greater depth of ground regions, due to perspective effect, the image size of a target varies significantly depending on the distance between the camera and the target. In this study, we propose a framework to transform high resolution images into warped images using a plane homography to make the target size uniform regardless of the position. The method not only reduces the number of pixels to be processed for speed-up, but also improves the tracking performance. We provide experimental results on object tracking in high-resolution maritime videos to demonstrate the validity of our method.

## 1. Introduction

The emergence of inexpensive high-resolution video cameras (>5 megapixels) benefits the video surveillance in a variety of ways. It potentially allows multiple targets to be detected simultaneously over large areas and tracked continuously for longer distances. High-resolution video also allows for more accurate detection, classification and identification of objects. However, the sheer volume of high-resolution image data imposes tremendous load on memory, network capabilities, processing time, and accuracy of tracking. Hence, real-time processing of high-resolution video is a challenge. In related literature,

Park and Trivedi use homography to register multiple camera views into the map view space [5]. Zhao and Nevatia use a homography for displaying tracking trajectories on the map view [7]. However, none of the methods warp input images for faster processing or better performance.

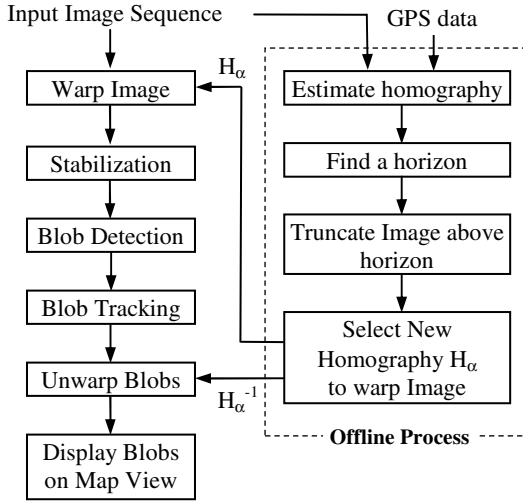


**Figure 1.** Image transformation using the proposed method. The transformation maps the rectangular image to a trapezoidal shape where the area near the camera view point is reduced so that the image size of a target remains relatively constant regardless of where it is located in the camera field-of-view.

In this work, we propose a novel method to speed up the processing time of object tracking and to reduce memory usage by warping images in the video stream, improving the accuracy of tracking. We focus on scenes where the ground plane is present and the surveillance camera is placed at a higher level than the targets, which is a common environment in surveillance applications. In this environment, the image size of a target near the horizon is much smaller than the one closer to the camera because of perspective effect. A small target may be treated as a noise in many detection methods. In contrast, the size of the target close to a camera is often more than that is required for proper detection and tracking. In addition, water reflection near a camera is severer than one in far-view, which prevents correct background modeling and object detection. We thus transform a rectangular image to a

trapezoidal shape where the area near the camera view point is reduced so that the image size of a target remains relatively constant regardless of where it is located in the camera field-of-view. With the reduced total number of pixels in the warped image, the system can process more frames in real time, which also improves object tracking performance. Figure 1 shows a 4000x1024 high-resolution image and its transformation.

This method consists of several main steps. First, we estimate the homography between the ground plane and the camera view using GPS data. Second, using the homography, the horizon is estimated and a warp area is determined. Third, a new homography which gives the best performance preserving tracking accuracies is introduced. Finally, each image in the video is warped based on the homography before tracking. Figure 2 shows the flowchart of our method.



**Figure 2.** Flowchart of the proposed method

The paper is organized as follows: Section 2 explains estimation and image truncation using the homography. Section 3 introduces a new warping method. Section 4 briefly explains detection and tracking methods. Section 5 shows the experimental results. Section 5 concludes the paper.

## 2. Image Truncation

The plane homography between a camera view and a ground plane is estimated using GPS data [3]. In case of land-based scene, a person carrying a GPS receiver and a wireless data transmitter walks across the field of view of a camera. In case of a maritime scene, a ship equipped with GPS receiver moves around both the near and far fields of a camera. The collected GPS data and corresponding manually annotated object trajectories on the images provide an accurate

estimation of homography  $\mathbf{H}$  from the ground plane (map view) to the camera view [4].

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

An inverse matrix  $\mathbf{H}^{-1}$  is a homography from the camera view to the map view. Using the homography, a horizon line  $\mathbf{V}$  can be easily extracted.

$$\mathbf{V} = \begin{bmatrix} h_{11} & h_{12} & x \\ h_{21} & h_{22} & y \\ h_{31} & h_{32} & 1 \end{bmatrix} = ax + by + c = 0$$

where  $a = h_{21}h_{32} - h_{31}h_{22}$ ,  $b = h_{12}h_{31} - h_{11}h_{32}$ ,  $c = h_{11}h_{22} - h_{12}h_{21}$ .

The width and height of the image are denoted as  $w$  and  $h$  respectively. The four boundary points can be represented by a matrix  $\mathbf{X}_b$  given by:

$$\mathbf{X}_b = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4] = \begin{bmatrix} 0 & w & w & 0 \\ 0 & 0 & h & h \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

As a first step of image reduction, we truncate the area above the horizon using a truncation matrix  $\mathbf{H}_T$  given by

$$\mathbf{H}_T = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & c/b \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where  $\theta = \tan^{-1}(-a/b)$ . The height and width are changed into  $h_N$  and  $w_N$  after transformation. If the horizon is out of the image or if the height of any target is higher than the height of a camera, we set  $\mathbf{H}_T$  as the identity matrix so that there is no truncation. In Figure 3-(a), an original image and the estimated horizon using the homography is shown. In Figure 3-(b), a truncated image after applying  $\mathbf{H}_T$  is shown.



(a) Original image with an estimated horizon (red line) using homography



(b) Truncated image below the horizon

**Figure 3.** Image Truncation

## 3. Image Warping

The second step of image reduction is accomplished by warping images using the map-to-view homography  $\mathbf{H}$ . An approach is to use  $\mathbf{H}$  directly for warping images which would give a better representation of the target

position on a rectified ground map. However, there are two problems in image warping using  $\mathbf{H}$ . The first is that the area near the horizon would be expanded to a very large area in the warped image while the area of interest near the camera, will be transformed to a very small area with low resolution. The second problem is that when a target is close to the horizon, the perspective distortion of the target becomes more severe, because the target is not an object lying flatly on the ground plane. As a result, the accuracy of the tracking decreases. Therefore, the map-to-camera view homography  $\mathbf{H}$  is not suitable for warping. Instead, a new homography between homography  $\mathbf{H}$  and truncation matrix  $\mathbf{H}_T$  is introduced. The new homography may transform a target to have less distortion and to have a relatively uniform size regardless of the position on the map.

The homography  $\mathbf{H}$  is modified to  $\mathbf{H}_M$  in order to let map-view images have the same resolution with camera-view images. First, the images of map-view are rotated and translated to be aligned to the x axis. Subsequently, the images are scaled to have the same width  $w_N$  and height  $h_N$  of the camera-view image.

$$\mathbf{H}_M = \mathbf{S} \cdot \mathbf{T} \cdot \mathbf{R} \cdot \mathbf{H}$$

$$= \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \rho & \sin \rho & 0 \\ -\sin \rho & \cos \rho & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{H} \quad (2)$$

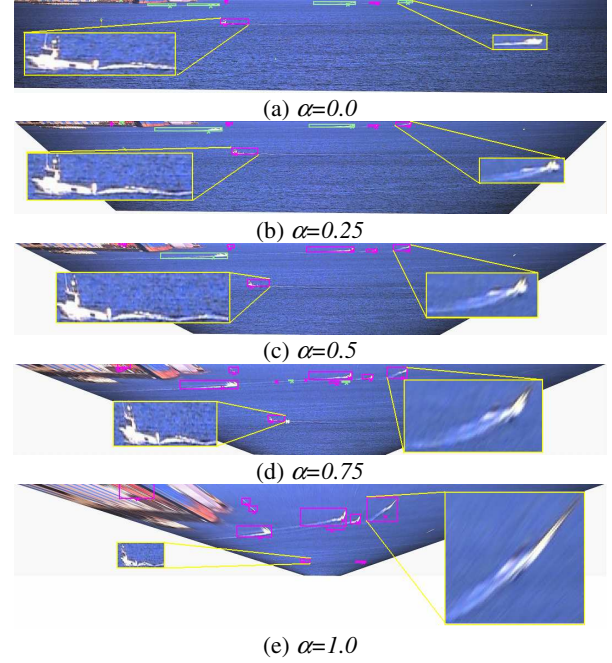
where  $\mathbf{R}$  is the matrix to rotate images according to x-axis,  $\mathbf{T}$  is the matrix moving the point  $\mathbf{x}_l$  to the origin.  $\mathbf{S}$  is the scaling matrix, and  $\rho = \pi/2 - \tan^{-1}(((\mathbf{H}\mathbf{x}_1)_x - (\mathbf{H}\mathbf{x}_2)_x) / ((\mathbf{H}\mathbf{x}_1)_y - (\mathbf{H}\mathbf{x}_2)_y))$ ,  $t_x = (\mathbf{R}\mathbf{H}\mathbf{x}_l)_x$ ,  $t_y = (\mathbf{R}\mathbf{H}\mathbf{x}_l)_y$ ,  $s_x = w_N / \|\mathbf{TRH}\mathbf{x}_1 - \mathbf{TRH}\mathbf{x}_2\|$ ,  $s_y = h_N / \max((\mathbf{TRH}\mathbf{x}_3)_y, (\mathbf{TRH}\mathbf{x}_4)_y)$ .  $(\cdot)_x$  and  $(\cdot)_y$  represents x and y value of a 2-D point after normalization of homogeneous coordinates.

The new homography  $\mathbf{H}_\alpha$ , which warps images between a camera view and a map view, is defined as

$$\mathbf{H}_\alpha = \alpha \mathbf{H}_M + (1 - \alpha) \mathbf{H}_T \quad (3)$$

where  $\alpha$  ( $0 \leq \alpha \leq 1$ ) determines the degree of image warping.  $\mathbf{H}_M$  should be normalized by dividing all elements by the last element  $h_{33}$  before applying Equation (3). When  $\alpha=0$ , the image has no warping and only the area above the horizon is truncated. When  $\alpha=1$ , the image are transformed to the map view, which has lowest memory consumption but has significant perspective distortion. The parameter  $\alpha$  determines the trade-offs between speed-up and distortion of the target. Figure 4 shows the image warping with various  $\alpha$  values. The target in the middle of image becomes

smaller and the targets near the horizon become larger, as  $\alpha$  increases. The proper selection of  $\alpha$  can decide the trade-off between good tracking performance in the camera view, memory saving, and speed-up. The best  $\alpha$  value depends on many factors such as camera view points and types of targets. The upper and lower bound of the memory space is  $w_N \cdot h_N$  and  $w_N \cdot h_N / 2$ .



**Figure 4.** Warped images with respect to  $\alpha$  and their tracking results for the same frame. Some tracked blobs are zoomed with the same ratio. When  $\alpha=0$ , the targets near a horizon are barely seen. As  $\alpha$  increases, the sizes of the targets near a horizon become bigger but the size of the target in the middle of the image becomes smaller.

## 4. Tracking

In this section, we briefly explain the tracking of a target as illustrated in Figure 2. All input images are warped using the homography  $\mathbf{H}_\alpha$ . Subsequently, images are stabilized using the KLT tracker [6] and RANSAC [1]. We estimate a translational motion to compensate for camera jittering. A moving object is detected as a blob using background subtraction, and it is tracked using the Kalman filter [7]. The position of the target is represented by its “footprint” which is the middle of the bottom edge of the bounding box. The footprint is a reliable feature in this transformation since it is on the ground plane. After the tracking, the footprint position of the target is transformed using  $\mathbf{H}^1 \cdot \mathbf{H}_\alpha^{-1}$  to obtain the position on the map view.

## 5. Experimental Results

We tested our algorithm on three sequences of maritime scenes. One of the videos is shown in Figure 1. The video is challenging due to various factors. The background is changing constantly due to waves and water reflections; fast moving watercraft produce long and wide wakes. In addition, strong wind causes severe camera jitter. However, the maritime video fits for the application of the proposed method since the ground plane is visible, the height of the camera is generally higher than the watercraft, and the video has considerable perspective effect. Each sequence has around 700 frames with  $4000 \times 1024$  pixel resolution. The footprints of all moving targets are manually annotated for evaluation.

The videos are processed with respect to varying  $\alpha$ . For each video, a homography  $\mathbf{H}$  is estimated and then a horizon is estimated. Consecutively, images are warped based on  $\mathbf{H}_\alpha$  using bilinear interpolation, and targets are detected and tracked. After tracking, the tracked targets are transformed to original positions using  $\mathbf{H}_\alpha^{-1}$ . We simulated real-time streaming of videos and measured the throughput of the system in terms of frame-per-second. The results were obtained on a computer with 2.8GHz Intel Core 2 Duo Extreme processor. Recall-Precision curves of footprint for the performance (Figure 5) and processing times are computed (Table 1) by changing  $\alpha$ . The recall and precision measures are computed using the distance between the “footprints” of the detected targets and the groundtruth annotation. This footprint distance is modulated by a shifted sigmoid function, given by:

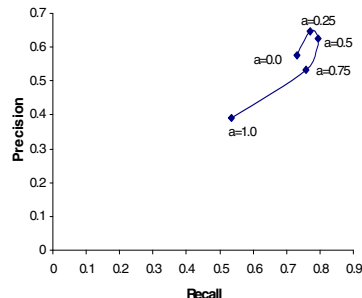
$$y = \frac{1}{1 - e^{\lambda(d-d_0)}}, \quad (4)$$

where  $y$  is a matching measure,  $d$  is the footprint distance (in pixels), and  $\lambda$  and  $d_0$  are the sigmoid parameters. In our evaluation, we used  $\lambda=0.5$  and  $d_0=15$  pixels. As  $\alpha$  was increased, the performance improved since the small targets near a horizon became bigger and the system could process more frames per second. However, as  $\alpha$  value approached 1, the performance deteriorated since objects were more distorted and the area near a camera had lower resolution where most objects passed by.

The average processing time of the proposed method with  $\alpha=0$  (3.7fps) had less computational gain compared to the original tracking method without applying warping (3.5fps) because our method performs additional computation for image warping. Best performance is obtained with  $\alpha=0.5$ , with a throughput of 5.3 fps (51.28% improvement).

**Table 1.** Average memory save (%) relative to the original image and processing throughput in frame per second (fps)

	Original	$\alpha=0$	$\alpha=0.25$	$\alpha=0.5$	$\alpha=0.75$	$\alpha=1$
Memory	100%	60.1%	50.6%	43.3%	37.6%	33%
Throughput	3.5 fps	3.7 fps	4.5 fps	5.3 fps	5.4 fps	5.5 fps



**Figure 5.** Plot of precision-recall measurement with varying  $\alpha$

## 6. Conclusion

Real-time object tracking with high resolution videos is a challenging task. In this paper, we proposed a simple but efficient image warping method to speed up the processing time with better performance by regularizing the object size and obtaining faster framerate. We tested the proposed method on maritime scenes and validated the performance. The method can be applied to many other scenes for object tracking.

**Acknowledgment:** This research was supported by the Office of Naval Research under Contract# N00014-08-C-84.

## References

- [1] M.A. Fischler and R.C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Comm. of ACM*, 24(6):381-395,1981.
- [2] F. Guo and R. Chellappa, “Video Mensuration Using a Stationary Camera,” *ECCV* (3) 2006: 164-176.
- [3] N. Haering, O. Javed, Z. Rasheed, K.H. Shafique, X. Cao, H. Liu, L. Yu, D. Madden, A. Chosak, G. Taylor, H. Gupta, A. Lipton, “Automatic Imaging Sensor Calibration using Objects that Broadcast Positional Information,” US Patent Application # 37112-230561.
- [4] R. Hartley and A. Zisserman, *Multiview Geometry in Computer Vision*, Cambridge Univ. Press, March 2004.
- [5] S. Park and M. Trivedi, “Homography-based Analysis of People and Vehicle Activities in Crowded Scenes,” *IEEE workshop on App. of Computer Vision*, 2007.
- [6] C. Tomasi, T. Kanade, “Detection and Tracking of Point Features,” *Technical Report CMU-CS-91-132*, Carnegie Mellon University, 1991.
- [7] T. Zhao, R. Nevatia, “Tracking Multiple Humans in Complex Situation,” *PAMI*, Vol. 26, No. 9, 1208-1221, September 2004.